Office de la propriété intellectuelle du Canada
Un organisme d'Industrie Canada

Canadian Intellectual Property Office
An Agency of Industry Canada

Bureau canadien des brevets

Certification

La présente atteste que les documents ci-joints, dont la liste figure ci-dessous, sont des copies authentiques des documents déposés au Bureau des brevets.

Canadian Patent Office

Certification

This is to certify that the documents attached hereto and identified below are true copies of the documents on file in the Patent Office.

Specification and Drawings, as originally filed, with Application for Patent Serial No: **2,293,920,** on December 31, 1999, by **NORTEL NETWORKS CORPORATION,** assignee of Maged E. Beshai, for "Global Distributed Switch".

Agent certificateur/Certifying Officer

December 1, 2000
Date

Canada

(CIPO 68)

OPIC CIPO

ABSTRACT OF THE DISCLOSURE

A flexible global distributed switch adapted for wide geographical coverage with an end-to-end capacity that scales to several Petabits per second (Pb/s), while providing grade-of-service and quality-of-service control is constructed from packet-switching edge modules and channel-switching core modules. The global distributed switch may be used to form a global Internet. The global distributed switch enables simple controls, resulting in scalability and performance advantages due to a significant reduction in the mean number of hops in a path between two edge modules. Traffic is sorted at each ingress edge module according to egress edge module. At least one packet queue is dedicated to each egress edge module. Harmonious reconfiguration of edge modules and core modules is realized by time counter co-ordination. The global distributed switch can grow from an initial capacity of a few Terabits per second to a capacity of several Petabits per second, and from regional to global coverage. It can accommodate legacy systems such as IP-based networks and provide clean connections among distant legacy devices, such as routers.

GLOBAL DISTRIBUTED SWITCH

TECHNICAL FIELD

This invention relates generally to the field of high capacity, wide area distributed data packet switching. In particular, the invention relates to an architecture for a global distributed switch constructed using a plurality of geographically distributed regional data packet switches.

BACKGROUND OF THE INVENTION

The explosive growth of the Internet and a corresponding growth in corporate communications have strained existing telecommunications infrastructure to an extent that connections may be blocked and quality of service may be degraded. Much of the poor performance of current networks can be attributed to the structure of the networks. In general, modern networks consist of a plurality of small capacity nodes interconnected by a plurality of links. Consequently, most connections require a plurality of "hops", each hop traversing a link between two nodes. It is well understood that as the number of hops involved in a connection increases, the more complex connection routing and control becomes, and the more quality of service is likely to be degraded. A high quality of service cannot be easily realized in a network of low capacity switches where a connection may require several hops, causing cumulative degradation of service quality.

It is well known that high capacity networks can reduce connection blocking and improve quality of service. In general, high capacity variable-size data packet switches, hereinafter referred to as universal switches, are desirable building blocks for constructing high performance, high capacity networks. A universal switch transfers variable-size packets without the need for fragmentation of packets at ingress. It is also rate regulated to permit selectable transport capacities on links connected to other universal switches. A universal switch is described in Applicant's co-pending United States Patent Application entitled RATE-CONTROLLED MULTI-CLASS HIGH-CAPACITY PACKET SWITCH which was filed on February 4, 1999 and assigned Serial No. 09/244,824, the specification of which is incorporated herein by reference.

Due to the high-volatility of data traffic in large networks such as the Internet and the difficulties in short-term engineering of such network facilities, a packet switch with an agile core is desirable. Such a switch is described in Applicant's co-pending United States Patent Application entitled SELF-CONFIGURING DISTRIBUTED SWITCH which was filed on April 6, 1999 and assigned Serial No. 09/286,431, the specification of which is incorporated herein by reference. In a switch with an agile core, core capacity allocations are adapted in response to variations in spatial traffic distributions of data traffic switched through the core. This requires careful co-ordination of the packet switching function at edge modules and a channel switching function in the core of the switch. Nonetheless, each edge module need only be aware of the available capacity to each other edge module in order to schedule

packets. This greatly simplifies the traffic control function and facilitates quality of service control.

Several architectural alternatives can be devised to construct an edge-controlled wide-coverage high capacity network. In general the alternatives fall into static-core and adaptive-core categories.

Static-core

In a static core switch, the inter-module channel connectivity is fixed (i.e., time-invariant) and the reserved path capacity is controlled entirely at the edges by electronic switching, at any desired level of granularity. Several parallel paths may be established between an ingress module and an egress module. The possible use of a multiplicity of diverse paths through intermediate modules between the ingress module and the egress module may be dictated by the fixed inter-module connectivity. A path from an ingress module to an egress module is established either directly, or through switching at an intermediate module. A connection, however, is controlled entirely by the ingress and egress modules. The capacity of a path is modified slowly, for example in intervals of thousand-multiples of a mean packet duration; in a 10 Gb/s medium, the duration of a 1 K-bit packet is a 100 nanoseconds while a path capacity may be modified at intervals of 10 milliseconds. An edge-controlled switch with a static core can grow to some 160 Terabits per second with the current state of the art in switching. The main impediment to network growth beyond this capacity is the fixed time-invariant channel connectivity in the core,

which may necessitate the use of multiple intermediate electronic switching points.

Adaptive-core

Control at the edge provides one degree of freedom. Adaptive control of core channel connectivity adds a second degree of freedom. The use of a static channel interconnection has the advantage of simplicity but it may lead to the use of long alternate routes between source and egress modules, with each alternate route switching at an intermediate node. The need for intermediate packet-switching nodes can be reduced significantly, or even eliminated, by channel switching in the core, yielding a time-variant, inter-modular channel connectivity. The channels are switched slowly, at a rate that is several orders of magnitude lower than the packet transfer rates.

In a vast switch employing an optical core, it may not be possible to provide a direct path of adaptive capacity for all module pairs. The reason is twofold: (1) the granularity forces rounding up to an integer number of channels and (2) the control delay and propagation delay preclude instant response to spatial traffic variation. However, by appropriate adaptive control of channel connectivity in response to variations in traffic loads, most of the traffic can be transferred directly with only an insignificant proportion of the traffic transferred through an intermediate packet switch.

There is a need, therefore, for a distributed switch for global coverage that enables end-to-end connections having a small number of hops, preferably not

exceeding two hops, and which is capable of adapting its core capacity according to variations in traffic loads.

Large, high-capacity centralized switches could form building blocks for a high-speed Internet. However, the use of a centralized switch would increase the access costs. There is therefore a need for a distributed switch that places the edge modules in the vicinity of traffic sources and traffic sinks.

OBJECTS OF THE INVENTION

It is therefore an object of the invention to provide a distributed switch that places the edge modules in the vicinity of traffic sources and traffic sinks.

It is a further object of the invention to provide a switch with an adaptive core that operates to provide sufficient core capacity in a shortest connection between each ingress edge module/egress edge module pair in the distributed switch.

It is also an object of the invention to provide a distributed switch with an optical core that can be reconfigured without disrupting the transfer of data at the ingress edge modules.

SUMMARY OF THE INVENTION

The invention therefore provides a high capacity distributed packet switch comprising a plurality of edge modules, each edge module including at least three input/output ports, the at least three input/output ports being organized in groups of J, K, and L input/output

ports. The J group of input/output ports is connected by communication links to a regional channel-switching core. The L group of input/output ports is connected by communications links to a global channel-switching core. The K input/output group of ports is connected by communications links to data traffic sources and data traffic sinks.

Edge modules having moderate capacities, 2 Tb/s each for example, can be used to construct a network of a Pb/s (Peta b/s) capacity if two-hop connections are acceptable for a significant proportion of the traffic. In a two-hop connection, packet-switching occurs at an intermediate edge module between an ingress edge module and an egress edge module.

The edge modules are preferably universal switches described in Applicant's co-pending Patent application filed February 4, 1999. A distributed packet switch of global coverage comprising a number of electronic universal switch modules interconnected by a distributed optical core is preferred. The distributed core comprises a number of memoryless core modules, and each core module comprises several parallel optical space switches. In order to enable direct connections for traffic streams of arbitrary rates, the inter-module connection pattern is changed in response to fluctuations in data traffic loads.

The capacity of a distributed switch in accordance with the invention is determined by the capacity of each edge module and the capacity of each of the parallel space switches in the core. The distributed switch enables an

economical, scalable high-capacity, high-performance Internet.

The distributed switch may be viewed as a single global switch having intelligent edge modules grouped into a number of regional distributed switches, the regional switches being interconnected to form the global switch. Although there is an apparent "hierarchy" in the structure of the global distributed switch in accordance with the invention, the global distributed switch in accordance with the invention is in fact a single-level, edge-controlled, wide-coverage packet switch.

BRIEF DESCRIPTION OF THE DRAWINGS

Further features and advantages of the present invention will become apparent from the following detailed description, taken in combination with the appended drawings, in which:

FIG. 1 is a schematic diagram illustrating an architecture for a global distributed switch in accordance with the invention, in which an edge module is connected directly to a single global core module by multi-channel links;

FIG. 2 is a schematic diagram illustrating the architecture of a global distributed switch in accordance with the invention, in which multi-channels from an edge module are shuffled so that the edge module is connected to each of several global core modules by one or more channels;

FIG. 3 is a schematic diagram illustrating an architecture for a global network in accordance with the invention, in which a plurality of channels from an edge module connect to several global core modules through a cross-connector to permit adjustments for long term changes in spatial traffic distributions;

FIG. 4a schematically illustrates the "shuffling" of wavelength division multiplexed (WDM) channels between a high capacity edge module and a distributed core in a global network in accordance with the invention;

FIG. 4b schematically illustrates the cross-connection of WDM channels between a high capacity edge module and a distributed core in a global network in accordance with the invention;

FIG. 5 is a schematic diagram of an exemplary medium capacity global distributed switch, the structure of which is modeled after the structure of the global distributed switch shown in FIG. 1.;

FIG. 6 is a schematic diagram in which a multi-channel (L channel) link from each edge module to the global core is connected to a shuffler adapted to shuffle the channels before they reach the global core; and

FIG. 7 is a connection matrix showing the possible allocation of inter-regional channels when the global distributed switch is connected as shown in FIG. 1.

It will be noted that throughout the appended drawings, like features are identified by like reference numerals.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

A high-performance global distributed switch with a capacity of an order of several Petabits per second (Pbs) is provided by the invention. In accordance with the invention, the high-performance global distributed switch includes high capacity (Terabits per second) universal switches as edge modules, and an agile channel-switching core. Control of the data packet network is exercised from the edge modules.

As shown in FIG. 1, a global distributed switch 20 in accordance with the invention includes a plurality of edge modules 22 that are clustered to form distributed regional switches 24. The distributed regional switches 24 are interconnected by global core modules 28, which are preferably adaptive optical switches, as will be explained below in more detail. The edge modules 22 are preferably universal data packet switches. A universal data packet switch is an electronic switch that switches variable-size data packets under rate control. The universal switch handles rate regulated traffic streams as well as "best effort" unregulated traffic streams. For the latter, a universal switch allocates service rates internally based on the occupancy of packet queues. The universal data packet switch is described in Applicant's co-pending United States Patent application entitled RATE-CONTROLLED MULTI-CLASS HIGH-CAPACITY PACKET SWITCH which was filed on February 4, 1999 and assigned Serial No. 09/244,824, the specification of which is incorporated herein by reference.

Distributed Regional Switch

As mentioned earlier, the global distributed switch is preferably constructed by interconnecting a number of regional distributed switches. This is discussed in further detail below. The distributed regional switch 24 includes N > 1 edge modules 22 and a number, C, of core modules. Each core module includes a number of space switches (not shown). A regional switching core 28 having J parallel space switches may be divided into a number of regional core modules that may be geographically distributed over an area bounded by a propagation-delay upper-bound (about 10 milliseconds). A channel-switching core having a number of channel-switching modules is described in Applicant's co-pending United States Patent Application Serial No. _____, filed on December 30, 1999 and entitled AGILE OPTICAL-CORE DISTRIBUTED PACKET SWITCH. In the description below, a regional switching core 28 may also be termed a regional core module.

Each space switch has N input ports and N output ports. The total number, J, of space switches in a regional channel-switching core 28 equals the number of inner channels, J, carried on links 30 connected to each edge module 22.

The regional core modules within a regional core 28 may have unequal sizes. The space switches in each regional core 28 are, however, identical and each regional core module includes at least two space switches. The non-uniformity in the size of the regional core modules sizes may be dictated by the spatial traffic distribution.

Regardless of the core type, optical or electronic, The channel-switching core is preferably partitioned for two reasons: economics and security.

An edge module is selected to host a core-module controller in addition to its function as an edge module, and is collocated and associated with each regional core module.

In a high capacity global network, each regional core switch 24 would include a number of distributed regional core modules, and each global core switch would include a number of distributed global core modules. Each core module, whether regional or global, must have an associated edge module for control purposes. In turn, each edge module must have a number of timing circuits equal to the number of core modules; one timing circuit dedicated to each core module (regional or global).

The edge modules and core modules are typically distributed over a wide area and the number of edge modules is much larger than the number of core modules. Each core module is associated with a selected edge module. The selected edge module is collocated with the associated core module and hosts the core module's controller. The association of a selected edge module with the channel-switching core is explained in Applicant's co-pending United States Patent Application Serial No. 09/286,431.

Distributed Regional Switch with an Optical Core

Each edge module has a fixed number W of one-way channels to the core, and it receives a fixed number, preferably equal to W, of one-way channels from the core.

The former are hereafter called A-channels, and the latter are called B-channels. A path from a module X to a module Y is formed by joining an A-channel emanating from module X to a B-channel terminating on module Y. Connecting the A-channel to the B-channel takes place at a core space switch. The number of paths from any module to any other module can vary from zero to W. The process of changing the number of paths between two modules is a reconfiguration process that changes the connection pattern of module pairs. A route from an edge module X to another edge module Y may have one path or two concatenated paths joined at an edge module other than modules X or Y. This is referenced as a loop path. A larger number of concatenated paths may be used to form a route. However, this leads to undesirable control complexity.

If the core is not reconfigured to follow the spatial and temporal traffic variations, a high traffic load from a module X to a module Y may have to use parallel loop-path routes. A loop-path route may not be economical since it uses more transmission facilities and an extra step of data switching at an edge module. In addition, tandem packet switching in the loop path adds to delay jitter.

It is emphasized that the objective of reconfiguration is to maximize the proportion of the inter-module traffic that can be routed directly without recourse to tandem switching in a loop path. However, connections from a module X to a module Y, which collectively require a capacity that is much smaller than a channel capacity preferably use loop-path routes. Establishing a direct

path in this case is wasteful unless the path can be quickly established and released, which may not be feasible. For example, a set of connections from an edge module X to an edge module Y collectively requiring a 100 Mb/s capacity in a switch core with a channel capacity of 10 Gb/s uses only 1% of a path capacity. If a core reconfiguration is performed every millisecond, the connection from edge module X to edge module Y could be re-established every 100 milliseconds to yield a 100 Mb/s connection. This means that some traffic units arriving at module X may have to wait for 100 milliseconds before being sent to module Y. A delay of that magnitude is unacceptable and a better solution is to use a loop path where the data traffic for the connections flows steadily via a tandem switched loop path through one of the edge modules other than edge modules X or Y.

Preferably, a regional distributed switch 24 is tolerant to core latency as described in Applicant's co-pending United States Patent Application Serial No. _____, filed on December 30, 1999 and entitled AGILE OPTICAL-CORE DISTRIBUTED PACKET SWITCH.

Distributed Regional Switch with an Electronic Core

In principle, the control of the data-transfer among the edge modules can be performed at the packet level. However, the high-rate of packet transfer may render the global controller unrealizable. For example, in a 100 Tb/s packet switch, at a mean packet length of 2000 bits, the packet arrival rate at full occupancy would be of the order of 50 Giga packets per second. Packet

transfer from ingress to egress under rate control may be scheduled periodically.

An alternative to packet-switching in the core is to establish inter-module paths of flexible capacities that may be allocated in sufficiently-large units to reduce the control burden by replacing the packet scheduler with a capacity scheduler. For example, if the inter-module capacity is defined in channel slots of 1/16 of the channel capacity, the total number of channel slots in a 100 Tb/s switch with 10 Gb/s ports would be about 160,000. The packet switch would reconfigure periodically, every 10 milli-seconds for example, or as the need arises in response to significant traffic-pattern changes. The time between successive reconfigurations is dictated by a propagation delay between the edge modules and the core modules, as will be discussed below. A capacity-scheduler computational load is thus lower than a core packet-scheduler load by three orders of magnitude. Preferably, a direct connection is provided for each edge module pair. The minimum capacity for a connection is a channel-slot capacity.

In order to provide direct paths for all edge-module pairs, an internal capacity expansion is required to offset the effect of under-utilized channels. The expansion may be determined by considering the case of full load under extreme traffic imbalance. Consider an extreme case where an edge module may send most of its traffic to another edge module while sending insignificant, but non-zero, traffic to the remaining (N-2) edge modules, resulting in (N-2) almost unutilized channel slots

emanating from the edge module. The maximum relative waste in this case is (N-2) / (S x M), where N is the number of edge modules, M is the number of channels per edge module, and S is the number of time slots per channel. With N = 256, M = 128, and S = 16, yielding a switch capacity of 320 Tb/s at a 10 Gb/s link capacity, the maximum relative waste is about 0.125. If N = 128, yielding a switch capacity of 160 Tb/s at a 10 Gb/s link capacity, the maximum relative waste is 0.0625.

It may be desirable, however, to aggregate traffic streams of low intensity in a conventional manner and perform an intermediate switching stage in order to avoid capacity waste. A traffic stream with an intensity of 0.1 of a channel-slot capacity can be switched at an intermediate point to conserve core capacity at the expense of a smaller waste in edge capacity. The threshold at which such a trade-off becomes beneficial is an engineering issue. Generally, it is desirable that only a very small proportion of the total traffic, preferably less than 5%, be switched at an intermediate point. This can be realized using a folded architecture. Alternatively, in an unfolded architecture, a common loop-back module may be shared for this purpose.

Global Switch

A global multi Peta-bits-per-second network, can be configured as shown schematically in FIG. 1. It includes a number of distributed regional switches 24, each with a capacity of the order of 40 to 160 Tb/s. The distributed regional switches 24 are interconnected by the global channel switches 26 shown on the right-side of FIG. 1. A

global channel switch 26 may be divided into a number of global core modules. The optical cross-connectors connecting the edge modules to the global channel switches are optional but their deployment adds a degree of freedom to the channel routing process resulting in increased efficiency.

Each electronic switch has three interfaces: a source/sink interface, a regional interface, and a global interface. A global core includes a number of optical cross connectors arranged in such a way as to render the network fully connected where each module can have a channel to each other module. The multiplicity of alternate paths for each switch-pair enhances the network's reliability.

Quadratic and Cubic Scalability

Two architectural alternatives can be used to realize a network of multi Peta bits per second capacity. The first uses edge modules of relatively-high capacities, of the order of 8 Tb/s each, for example, and the second uses edge modules of moderate capacities, of the order of 2 Tb/s each. The total capacity in the first architecture varies quadratically with the edge-switch capacity. The capacity in the second architecture is a cubic function of the edge-switch capacity. The merits of each of the architectures will be highlighted below.

Quadratic Scalability

An edge module has (K + L) dual ports comprising (K + L) input ports and (K + L) output ports. The K dual ports are connected to traffic sources and sinks. The L

dual ports are connected to a maximum of L other edge modules by channels of capacity R bits/second each, yielding a fully-meshed network of (L+1) edge modules. The maximum capacity is realized if each source edge module sends all its traffic to a single sink edge module, thus reducing a distributed switch of N dual edge modules (N source modules paired with N sink modules) to N point-to-point isolated connections. This trivial hypothetical case is, of course, of no practical interest. The maximum non-trivial traffic capacity of the fully-meshed network is realized when the traffic load is spatially balanced so that each edge module transfers the same traffic load to each other edge module. The network capacity is then $C = \eta \times K \times (L+1) \times R$, η being the permissible occupancy of each channel, all the edge to edge traffic loads being statistically identical. If the network is designed so that the capacities from each source module to each sink module are equal, then, with spatial traffic imbalance, a source edge module may have to transfer its traffic load to a given sink module through one or more other edge modules. The source and sink edge modules should therefore be paired and each pair should share the same internal switching fabric. The use of intermediate packet-switching modules results in reducing the meshed-network traffic capacity below the value of C. To accommodate violent traffic variations, the network should be designed for high traffic variation. In the extreme case where each edge module temporarily sends its entire traffic to a single sink module, other than its own conjugate sink module, the traffic capacity is slightly higher than $0.5 \times C$. If a proportion of the traffic emanating from each source module

is transferred through an intermediate packet-switching module, then the ratio L/K must be greater than 1.0 and the overall traffic efficiency is about K/L. The ratio K/L depends on the spatial traffic imbalance, and a mean value of 0.7 would be expected in a moderately-volatile environment. The capacity of an edge module, which equals $(L + K) \times R$, limits the network capacity. With a ratio of L/K of 1.4, an edge module having a total number of dual ports of 384 ports for example (384 input ports and 384 output ports), with R = 10 Gb/s and $\eta = 0.90$, yields a maximum network capacity of about 288 Tb/s (K = 160, L = 224). The ratio of the outer capacity to inner capacity is about 0.70.

The ratio of outer capacity to inner capacity increases with core agility. In the extreme case described above, this ratio is slightly higher than 0.5 in a static-core but can reach a value of about 0.95 with an agile self-configuring optical core.

If it is possible to adapt the core connections to traffic loads so that the capacities from a source edge module to a sink edge module is a function of the respective traffic load, then the overall capacity can be maximized to approach the ideal maximum capacity. In such case, the internal expansion (L/K) can be reduced and with the 384-port edge module, K and L may be chosen to be 184 and 200, respectively, yielding a regional distributed-switch capacity of 330 Tb/s.

Cubic Scalability

An edge module has (J + K + L) dual ports comprising (J + K + L) input ports and (J + K + L) output ports. The K dual ports are connected to traffic sources and sinks. The J dual ports are connected to a maximum of J other edge modules by a channel of capacity R bits/second each, yielding a fully-meshed network-region of (J+1) edge modules. The maximum traffic capacity of a regional distributed switch 24 being $c = \eta \times K \times (J+1) \times R$. The L dual ports are connected to L other network regions. With a static core, each source edge module is connected to a sink edge module in the same region by a direct channel or a maximum of J alternate paths each of a maximum of two hops, i.e., two channels in series. The total number of dual edge modules is then $(J+1) \times (L+1)$. With a static core, each source module can reach each sink module of a different network region through several alternate paths of at most two hops each. With a static core of uniform structure, only one sink module in a region is directly connected to a source edge module in another region. The maximum traffic capacity of the two-hop network is realized when the traffic load is spatially balanced so that each edge module transfers the same traffic load to each other edge module. The network capacity is then $C = \eta \times K \times (J+1) \times (L+1) \times R$, η being the permissible occupancy of each channel as defined above, all the edge to edge traffic loads being statistically identical. With the same edge-module parameters used above (384 dual ports each, R = 10 Gb/s, and η = 0.90), and selecting L = K = J = 128, the overall capacity grows to about 18.8 Pb/s. The ratio of the outer capacity to the inner capacity is 0.5.

Again, the objective of an agile core is to adapt the ingress/egress capacity to follow the traffic load.

FIG. 5 is a schematic diagram of an exemplary medium capacity distributed switch 20, the structure of which is modeled after the structure of the global distributed switch 20 shown FIG. 1. The configuration of the global distributed switch shown in FIG. 5 is small for simplicity of illustration. There are four regional switches 24, each having four edge modules 22. The four regional switches 24 have one regional core module 28 each. The regional core modules 28 are labeled RC0 to RC3. The edge modules associated with RC0 are labeled a_0 to a_3, the edge modules associated with RC1 are labeled b_0 to b_3, and so on. There are four global core modules 26 labeled GC0 to GC3 interconnected as shown in FIG. 5. In the architecture shown in FIG. 5, each edge module 22 connects to only one of the global core modules 26. For example, edge module a_0 connects by a two-way multi-channel link 32 to global core module GC0, while edge module A_1 connects by a two-way multi-channel link 32 to global core module GC1, and so on.

The connectivity of the distributed switch shown in FIG. 5 is indicated in connection matrix 100 shown in FIG. 7. An entries in the matrix marked 'x' indicates a direct path of one or more channels. The connection matrix 100 shows each region to be fully connected. An edge module can connect to any other edge module in the same region via a direct path of adjustable capacity. The interconnection between regions takes place through the diagonal of connectivity shown in blocks 102.

FIG. 6 is a schematic diagram of a configuration for a global distributed switch 20 in which a multi-channel (L channel) link from each edge module to the global core is connected first to a shuffler 40 or a cross-connector 42. The shuffler 40 is similar to the one shown in FIG. 4a, which shuffles 4-wavelength optical links. The shuffling of channels (wavelengths) results in enabling the inter-regional connectivity to be more controllable, thus increasing the opportunity to establish direct connections between an edge module in one region and an edge module in another region. In other words, this increases the proportion of single-hop connections, hence increasing the overall traffic capacity of the global distributed switch.

The connection matrix for the shuffler 40 shown in FIG. 6 is illustrated in FIG. 7. The allocation of the inter-regional channels is thus made more flexible.

The cross-connector 42 shown in FIG. 6 permits the inter-regional connectivity to be further enhanced by facilitating adaptive and unequal inter-regional channel assignment. This permits better adaptation to extreme and unpredictable spatial traffic variations. As will be understood by those skilled in the art, the multi-channel links are preferably optical wavelength division multiplexed (WDM) links.

Control

Each edge module should have a timing circuit dedicated to each regional or global core module. If a regional core 28 includes C1 regional core modules and the total number of global core modules is C2, then each edge

module 22 must have (C1 + C2) timing circuits. A detailed description of a preferred timing circuit is described in United States Patent Application Serial No. 09/286,431 filed April 6, 1999 and entitled SELF-CONFIGURING DISTRIBUTED SWITCH, the specification of which is incorporated by reference.

Time-counter Period

Using an 18-bit time counter with a 64 nano-second clock period yields a timing cycle of about 16 milliseconds. With a one-way propagation delay between an edge module and any core module, of the order of 10 milliseconds within a region, a time-counter period of 16 milliseconds is adequate.

A 22-bit global time counter yields a timing period of 256 milliseconds with a clock period of 64 nanoseconds (about 16 Mega Hertz). This timing period is adequate for global coverage.

Reconfiguration Rate

A regional core module should be reconfigured frequently to increase the agility of the regional distributed switch. Thus, it is preferable to define a network region according to geographic boundaries so that the propagation delay can be contained within acceptable bounds.

As described earlier, edge modules within a network region are interconnected by regional core modules to form a regional distributed switch. Several regional distributed switches are interconnected by global core modules to form a global network.

Global core modules can not reconfigure in short periods, for example within a 20 millisecond period, due to the large propagation delay. The one-way propagation delay between an edge module and a global core module can be of the order 100 milliseconds and the time alignment process described above requires an interchange of timing packets between reconfiguring edge modules and core modules. This requires that the reconfiguration period be larger than the propagation delay from a source edge module to any core module.

Special Case

In one extreme, the number of ports J can be selected to be zero, and each edge module connects only to core global modules, either directly, through a shuffle stage, or through a cross connector. In such a case, the reconfiguration rate is low, each 256 milliseconds for example.

The regional core modules should, preferably, have the same space-switch size, e.g., all using 32×32 space switches. However, the number of parallel space switches in a core module may differ from one core module to another. For example, 128 regional-interface ports (J = 128) may be divided into four regional core modules having 20, 24, 32, and 52 parallel space switches. The selection of the number of space switches per core module is governed by the spatial distribution of the source modules and their respective traffic intensity.

A space switch in a global core module is preferably of a higher capacity than that of a regional

core module. For example, while a regional core module may be of size 32×32, a global core module preferably uses 256×256 parallel space switches.

Long-term Configuration

A designated controller collects traffic data and trends them to characterize long-term traffic patterns. This data is used to determine the connectivity of the cross-connector 42 The rate of reconfiguration of cross-connectors is very slow with long intervals between successive changes. Prior to any change in a cross-connection pattern, a new route-set is computed offline and communicated to respective edge modules.

Mixture of Core Switches

The scalability of global network is determined by both the edge modules and core modules. The capacity of a regional switch is determined by the capacity of each of the parallel space switches. The latter determines the number of edge modules that can be interconnected through the regional core modules.

The number of regional switches that can be interconnected to form a 2-hop connected global network, where each edge module in a network region can reach any other edge module of another region in at most two hops, is determined by the capacity of each of the parallel space switches in a global core module.

An electronic space switch is appropriate for global core modules due to their scalability (to more than

256x256 for example, with each port supporting 10 Gb/s or more).

Different regional or core modules may use optical or electronic space switches. However, preferably, a specific core module should use the same type of space switches; optical or electronic.

Independent Master Timing vs. Globally-coordinated Timing

Each regional core module has its own controller which is supported by an edge module collocated with the regional core module. Similarly, each global core module has its own controller. The timing circuit of each core is independent of all other timing circuits. There is no benefit in using a global timing reference.

Internal Routing

Traffic is sorted at each source edge switch according to sink edge switch. At least one packet queue is dedicated to each sink edge module.

For brevity the term "source" is used to denote a source edge module and a "sink" denotes a sink edge module. For each source-sink pair (source: source edge module, sink: sink edge module), sets of single-hop, two-hop, and three-hop routes are determined. Some sets can be empty: for example, some source-sink pairs may not have direct (single-hop) paths. The process of determining the sets is straightforward. The sets may also be sorted according to cost, in which case a three-hop path may be perceived as less expensive than a two-hop path.

The main premise of an ideal agile-core network is that the traffic can always be routed through the shortest path, and the shortest path is continually adapted to have appropriate capacities for the load it is offered.

When the shortest path is fully assigned, a capacity increment is requested. If shortest-path capacity can not be increased, the connection may be routed on the second best path. If the latter can not accommodate the connection, a request is made to increase capacity of the second-best route. Finally, if steps (1) and (2) fail to accommodate the connection, an attempt is made to route the connection through the third-best path. Another option is to focus only on enhancing the capacities of the shortest paths and use alternate paths in an order of preference determined by their relative costs.

The embodiment(s) of the invention described above are intended to be exemplary only. The scope of the invention is therefore intended to be limited solely by the scope of the appended claims.

THE EMBODIMENTS OF THE INVENTION IN WHICH AN EXCLUSIVE PROPERTY OR PRIVILEGE IS CLAIMED ARE DEFINED AS FOLLOWS:

1. A high capacity distributed packet switch comprising:

   a) a plurality of edge modules, each edge module including at least three input/output ports, the at least three input/output ports being organized in groups of J, K, and L input/output ports; wherein

   b) the J group of input/output ports is connected by communication links to a regional channel-switching core;

   c) the L group of input/output ports is connected by communications links to a global channel-switching core; and

   d) the K group of input/output ports is connected by communications links to data traffic sources and data traffic sinks.

2. The high capacity distributed switch as claimed in claim 1 wherein the regional channel-switching core comprises a number of spatially distributed regional core modules and the global channel switching core comprises spatially distributed global core modules.

3. The high capacity distributed switch as claimed in claim 2 wherein the regional and global channel-switching cores comprise memoryless switches.

4. The high capacity distributed switch as claimed in claim 2 wherein a core module comprises a plurality of parallel optical space switches.

5. The high capacity distributed switch as claimed in claim 1 wherein the plurality of edge modules are divided into groups and the J dual-ports of each edge switch belonging to a one of the groups are connected exclusively to a respective regional channel-switching core.

6. The high capacity distributed switch as claimed in claim 1 wherein the input/output ports of the L group of each edge module in a group of edge modules are connected directly to only one of the global core modules.

7. The high capacity distributed switch as claimed in claim 6 wherein the input/output ports of the L group of two or more of the edge modules in a group of edge modules are respectively connected to two or more of the global core modules via a memoryless shuffle stage.

8. The high capacity distributed switch as claimed in claim 6 wherein the input/output ports of the L group of each of the edge modules in a group of the edge modules are respectively connected to two or more global core modules via a memoryless cross-connector.

9. The high capacity distributed switch as claimed in claim 2 wherein the regional core modules and their

associated edge modules are spatially separated in a geographical zone bounded by a distance at which a propagation-delay of signals traveling on the links between any core module and any associated edge module is within a predetermined upper bound.

10. The high capacity distributed switch as claimed in claim 1 wherein a path between any two successive edge modules in a route passes through at most one adaptive channel-switching stage.

11. The high capacity distributed switch as claimed in claim 2 wherein an edge module is collocated and associated with each regional core module and each global core module, and a core-controller is hosted by each of the edge modules collocated with the respective core modules.

12. The high capacity distributed switch as claimed in claim 1 wherein each edge module maintains a routeset to every other edge module in the distributed switch, the elements of each routeset identifying routes to a respective other edge module.

13. The high capacity distributed switch as claimed in claim 12 wherein the routes in each route-set are sorted according to a predetermined criterion.

14. The high capacity distributed switch as claimed in claim 2 wherein a regional or a global core module is adaptively reconfigured in response to fluctuations in data traffic loads.

15. The high capacity distributed switch as claimed in claim 1 wherein a cyclic time period of a control timing circuit of a regional core module is substantially shorter than a cyclic time period of a control timing circuit of a global core module.

16. The high capacity distributed switch as claimed in claim 15 wherein the control timing circuit for each of the regional core modules comprises an 18-bit counter, the control timing circuit for each of the global core modules is a 22-bit counter, and the clock rate for all of the regional and global core modules is 16 megahertz.

17. The high capacity distributed switch as claimed in claim 1 wherein a rate at which a global core module is reconfigured is substantially lower than a rate at which a regional core module is reconfigured.

18. The high capacity distributed switch as claimed in claim 1 wherein the communications links in the group L are optical links that support wavelength multiplexed data channels.

19. The high capacity distributed switch as claimed in claim 18 wherein the wavelength multiplexed data channels from the plurality of edge modules are shuffled together into a plurality of wavelength multiplexed links, each link carrying wavelengths from one or more of the edge modules.

20. The high capacity distributed switch as claimed in claim 18 wherein the wavelength multiplexed data channels from the plurality of edge modules are cross connected to a plurality of wavelength multiplexed links.

21. The high capacity distributed switch as claimed in claim 2 wherein the regional core modules have respective switching capacities, and the switching capacities are different.

22. The high capacity distributed switch as claimed in claim 2 wherein the global core modules have respective switching capacities, and the switching capacities are different.

23. The high capacity distributed switch as claimed in claim 2 wherein the regional core modules comprise optical space switches, electronic space switches or a combination of optical space switches and electronic space switches.

24. The high capacity distributed switch as claimed in claim 2 wherein the global core modules comprise optical space switches, electronic spaces switches or a combination of optical space switches and electronic space switches.

25. The high capacity distributed switch as claimed in claim 2 wherein all of the core modules comprise static cross-connectors.

26. The high capacity distributed switch as claimed in claim 1 wherein the value of J is zero and the edge modules interconnect solely with the global core modules.

27. The high capacity distributed switch as claimed in claim 8 wherein the memoryless cross-connects are configured based on long term spatial traffic distribution estimations or projections.

28. The high capacity distributed switch as claimed in claim 27 wherein new route-sets are distributed to each edge module controller prior to reconfiguration of the slow optical switches.

SWABEY OGILVY RENAULT
Suite 1600
1981 McGill College Avenue
Montreal, Quebec, Canada
H3A 2Y3

Patent Agents for the Applicant.

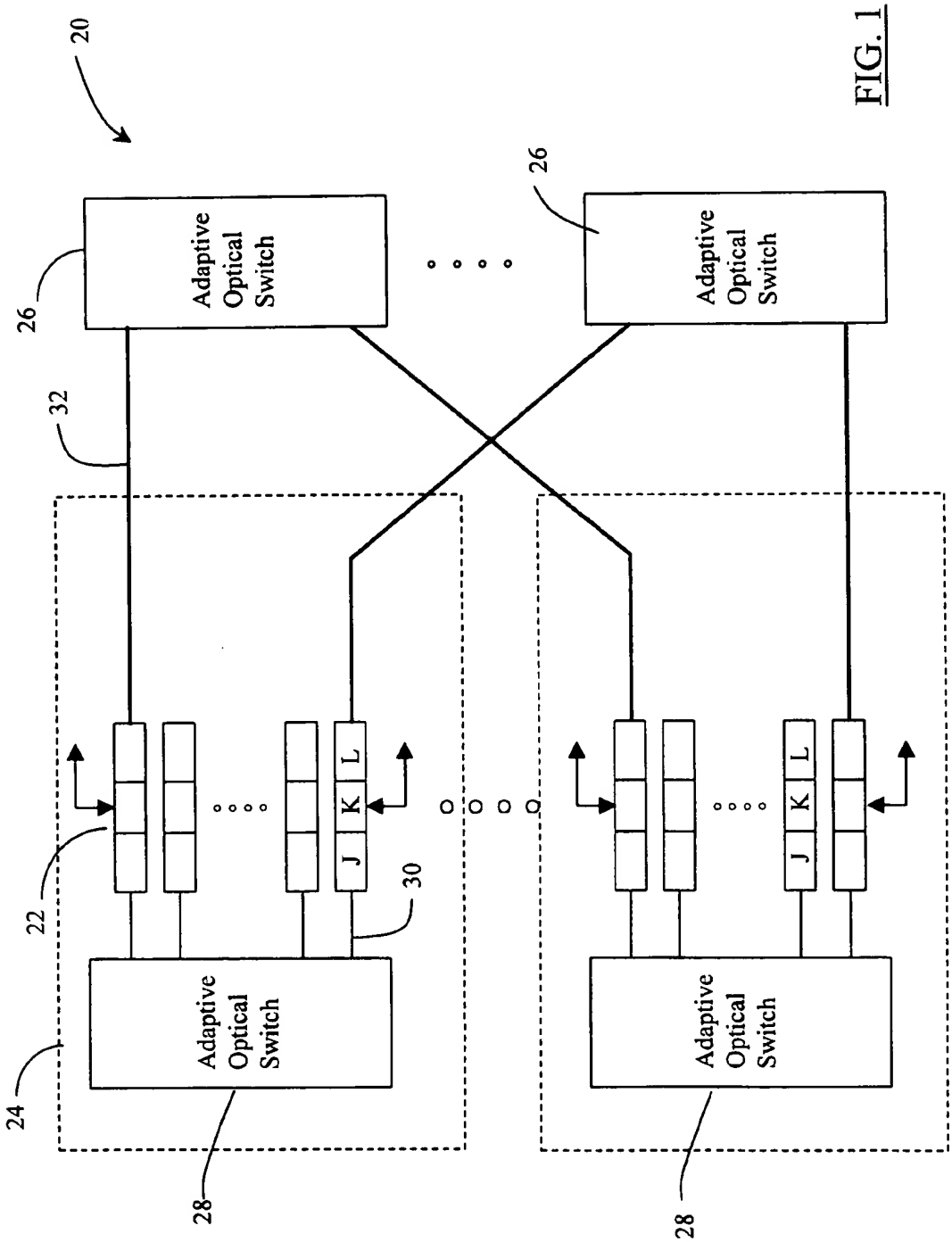
20
26
Adaptive Optical Switch
26
Adaptive Optical Switch
32
22
J
K
L
30
Adaptive Optical Switch
24
28
Adaptive Optical Switch
28

FIG. 1

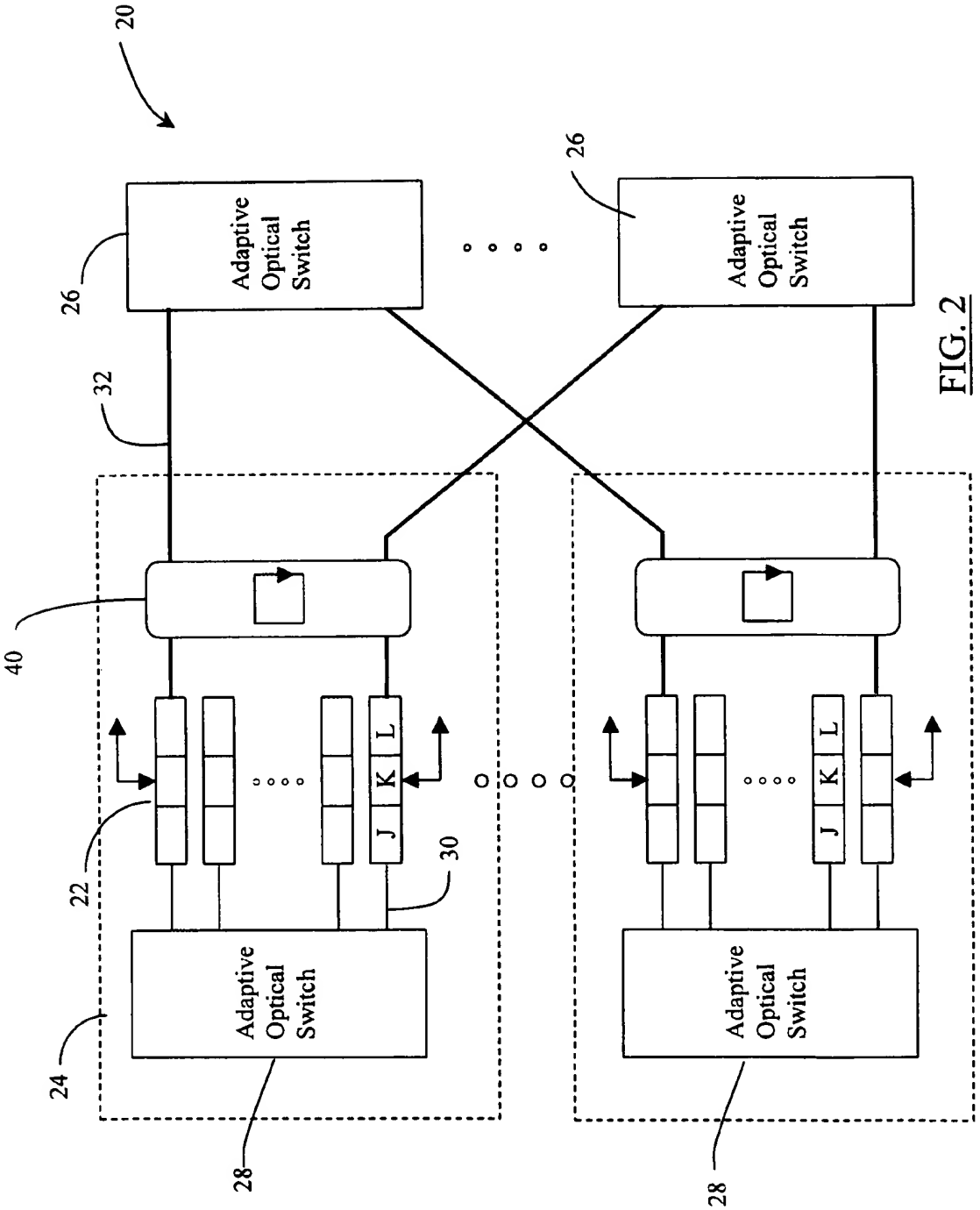
20
26
Adaptive Optical Switch
26
Adaptive Optical Switch
32
40
22
30
J
K
L
Adaptive Optical Switch
24
28
Adaptive Optical Switch
28

FIG. 2

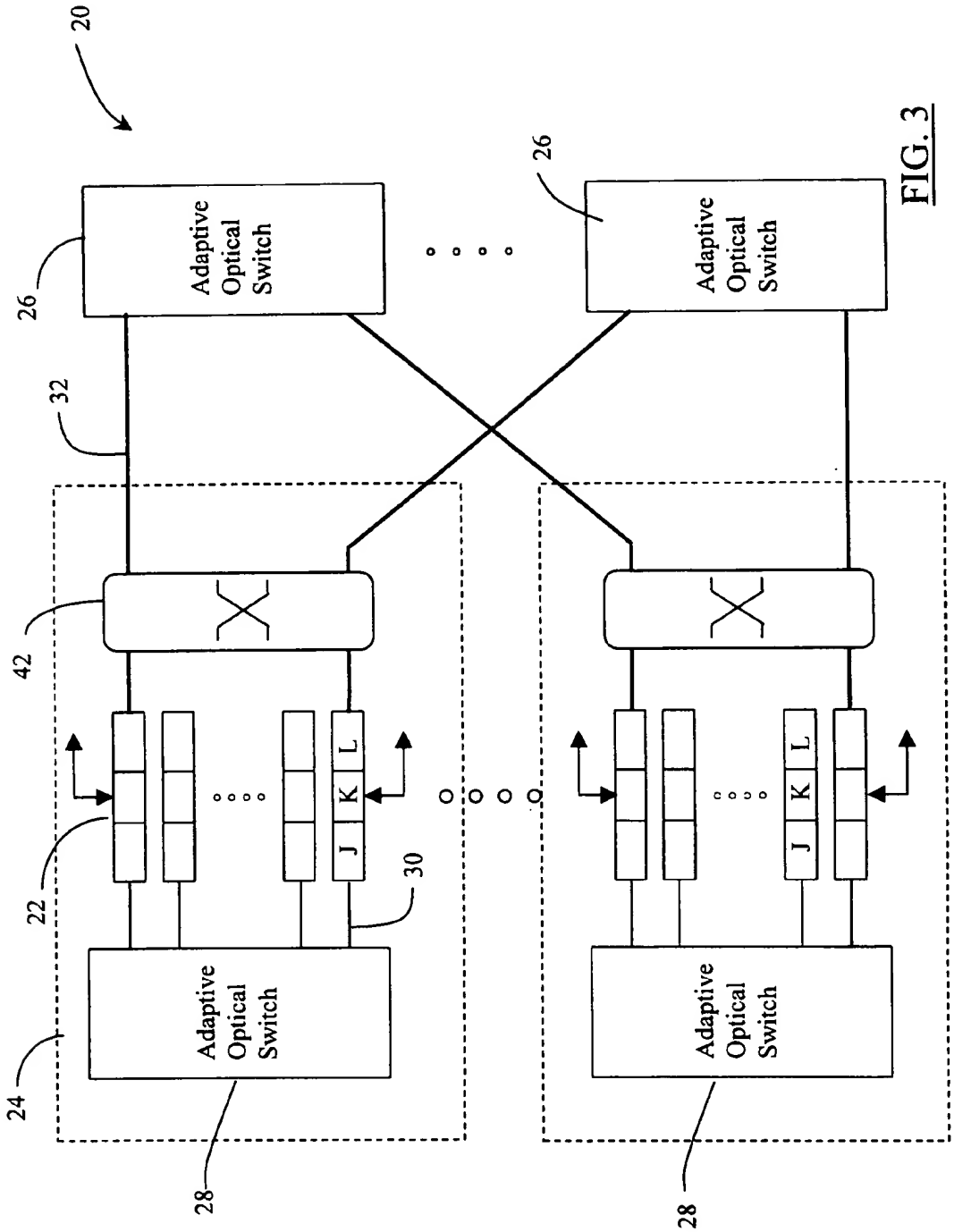
20
26
Adaptive Optical Switch
26
Adaptive Optical Switch
32
42
22
J
K
L
30
Adaptive Optical Switch
24
28
Adaptive Optical Switch
28

FIG. 3

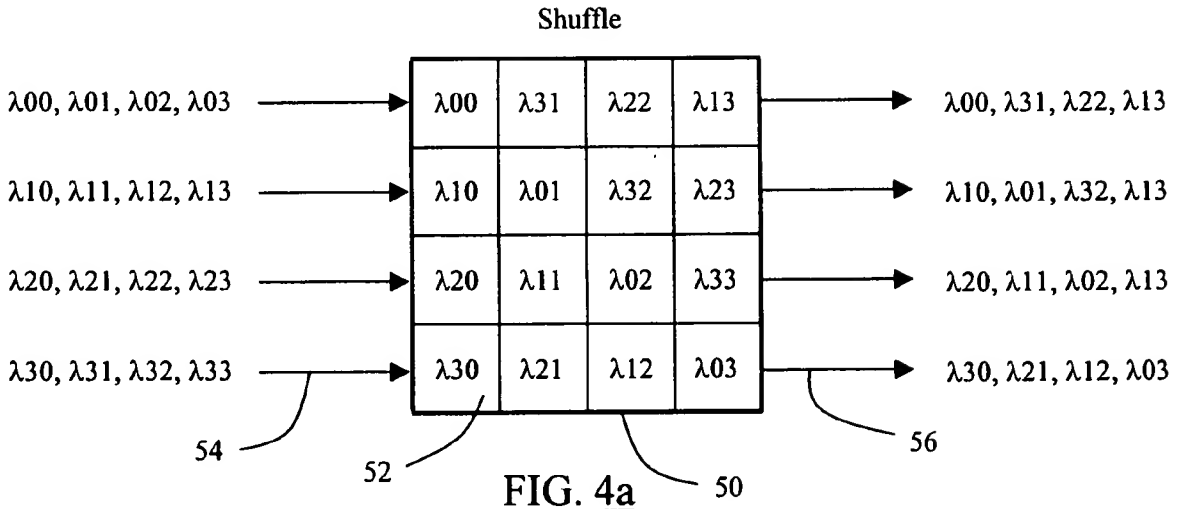

FIG. 4a
(Prior Art)

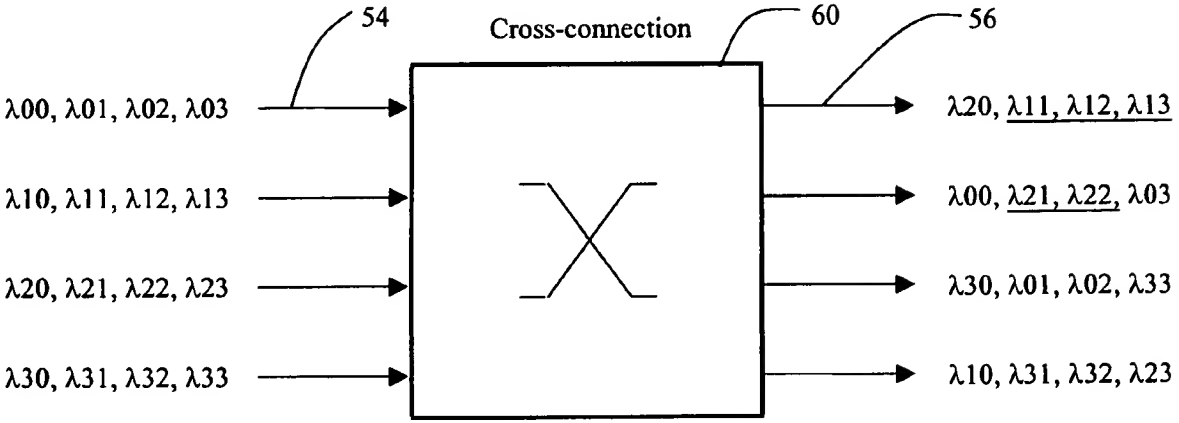

FIG. 4b
(Prior Art)

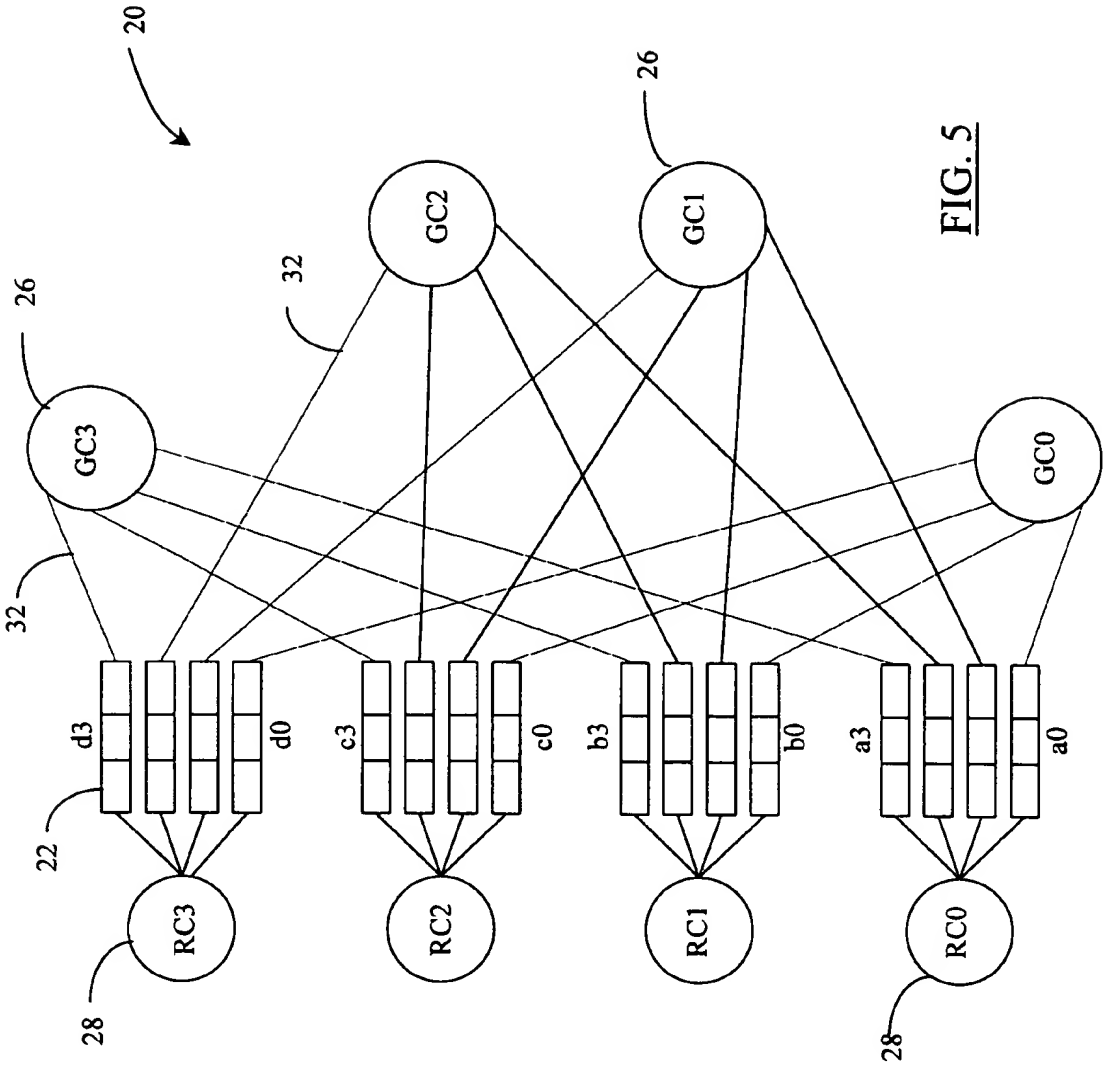
20
26
GC2
GC1
26
32
GC3
GC0
32
d3
d0
c3
c0
b3
b0
a3
a0
22
RC3
RC2
RC1
RC0
28
28

FIG. 5

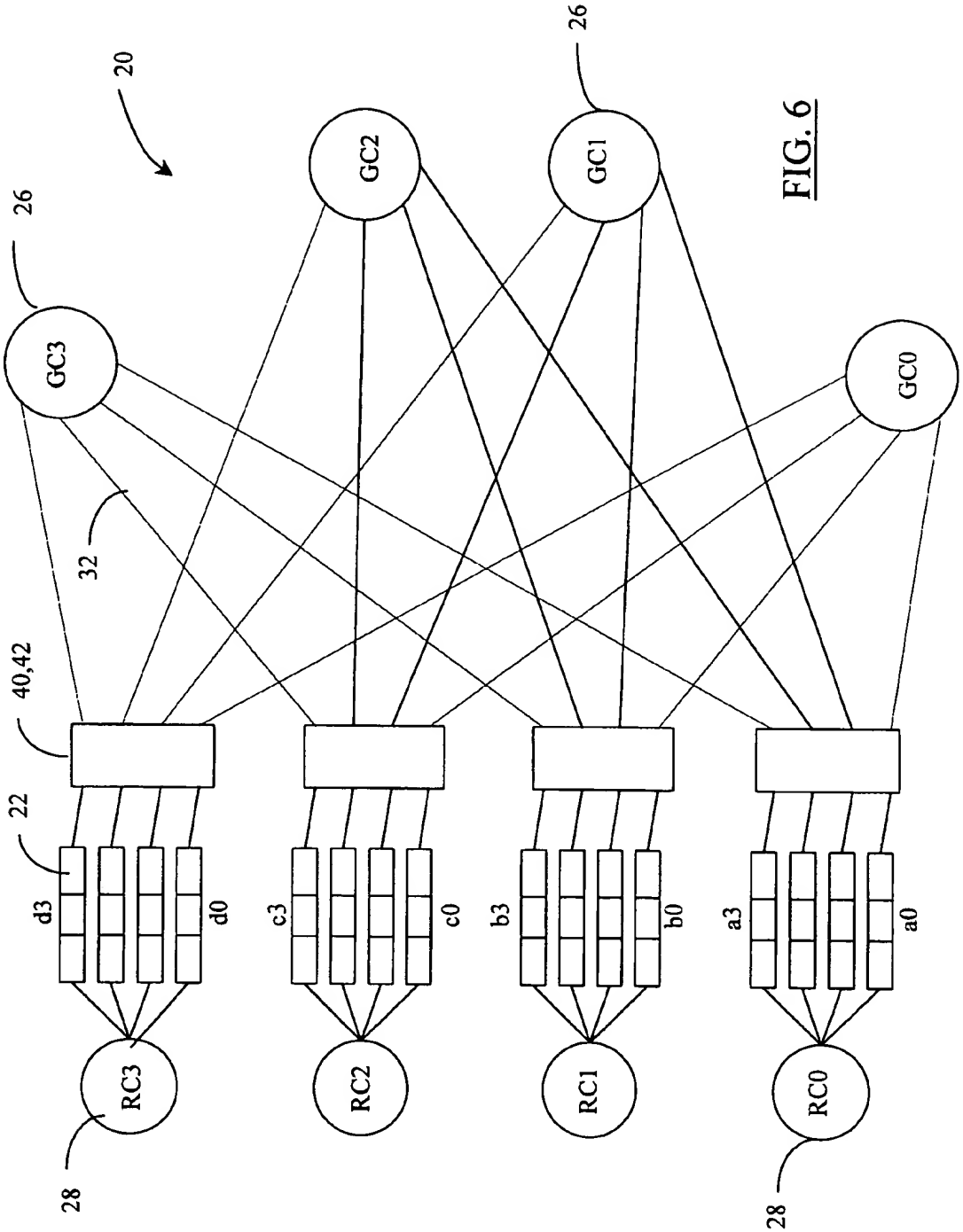
20
26
GC3
GC2
GC1
GC0
32
40,42
22
d3
d0
c3
c0
b3
b0
a3
a0
RC3
RC2
RC1
RC0
28
28

FIG. 6

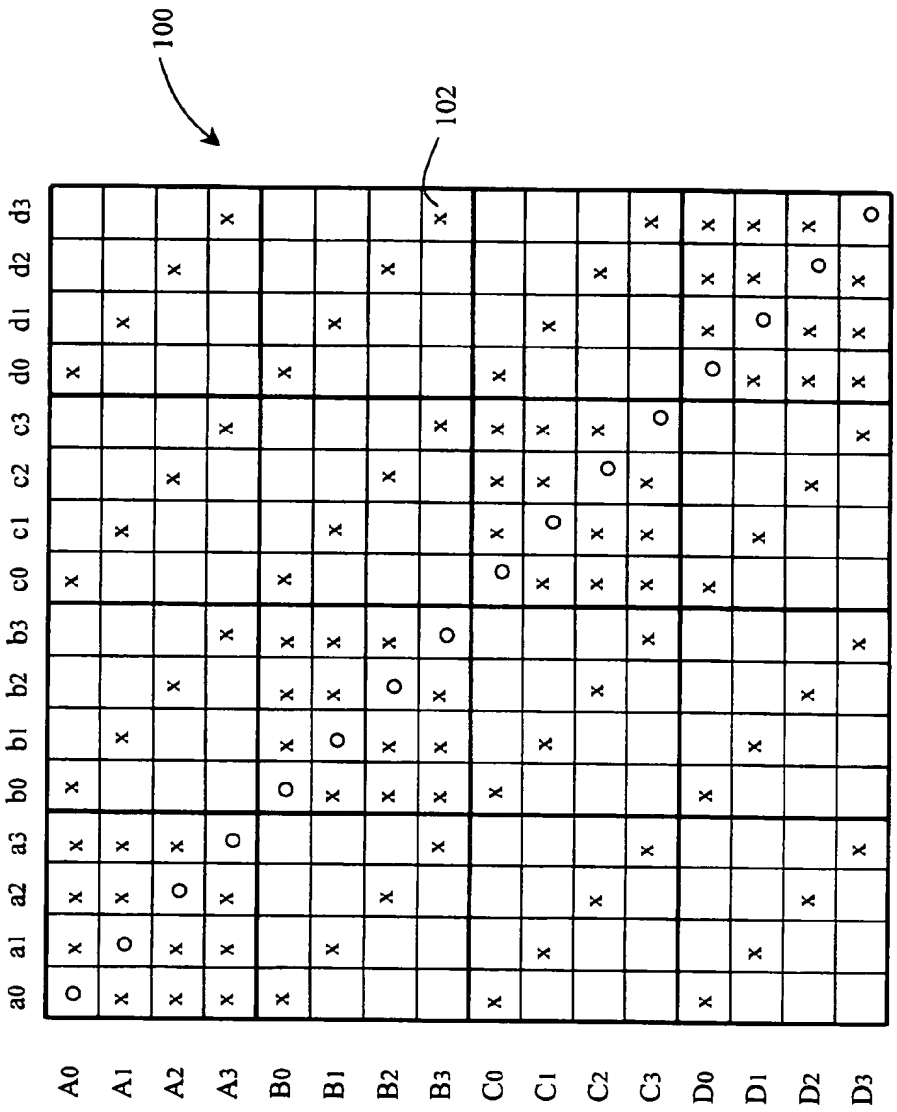
100
102
a0 a1 a2 a3 b0 b1 b2 b3 c0 c1 c2 c3 d0 d1 d2 d3
A0 A1 A2 A3 B0 B1 B2 B3 C0 C1 C2 C3 D0 D1 D2 D3

FIG. 7